UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Anotación del genoma humano y de ratón

Máster Universitario en Bioinformática y Biología Computacional

Autor: Martínez_Gómez, Laura

Tutor: Tress, Michael
Unidad de Bioinformática
Centro Nacional de Investigaciones Oncológicas

FECHA: febrero, 2018

ANOTACIÓN DEL GENOMA HUMANO Y DE RATÓN

AUTOR: Laura Martínez Gómez
TUTOR: Michael Tress
PONENTE: Luis Del Peso Ovalle

Centro Nacional de Investigaciones Oncológicas
Unidad de Bioinformática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Febrero 2018

ÍNDICE

RESUMEN

A pesar de que han pasado más de diez años de la publicación de la secuencia del genoma humano, sigue sin esclarecerse un número final de genes codificantes. Para el estudio y comprensión de la biología humana es fundamental estar en posesión de un catálogo de los genes codificantes contenidos en el genoma humano que sea preciso y que no contenga entradas de dudosa evidencia experimental, que sirva como base a la hora de hacer estudios biomédicos de gran escala. La combinación de diversos métodos computacionales de predicción de genes con la revisión exhaustiva de los grupos de anotadores manuales, nos ha acercado a una versión final del catálogo de genes humanos codificantes. Sin embargo, hay un porcentaje de genes anotados como codificantes provenientes de predicciones automáticas cuyo potencial codificante es dudoso. En este proyecto se ha desarrollado un pipeline automático que hace uso de 15 características provenientes de distintas bases de datos como GENCODE, Ensembl y UniProt, las anotaciones de estructura y función disponibles en la base de datos de *splicing* alternativo de APPRIS y la conservación entre especies medida por PhyloCSF y Compara. Del total de 20,292 genes codificantes provenientes de GENCODE, se han encontrado 2,467 genes que tienen al menos una de estas características que los convierte en potencialmente no codificantes. 66 de estos genes ya han sido eliminados por el equipo de anotación manual de HAVANA y alrededor de 30 han quedado para discusión.

Palabras Clave

Genes no codificantes, GENCODE, anotación humana

ABSTRACT

Despite its importance, the catalogue of human coding genes is not yet complete. In order to better understand the human genome sequence, it is crucial to avoid dubious protein coding entries within the genome annotation.

The combination of both different gene prediction computational methods and rigorous manual annotation has brought us considerably closer to a final catalogue of human coding genes. Nevertheless, there is a number of protein-coding annotated genes, those with the most conflicting evidence, which coding potential is still to be discussed.

Here we show that 12% of the coding genes in the Ensembl91/GENCODEv27 annotation of the reference catalogue have at least one potential non-coding feature. This result comes from the analysis of the data obtained from different annotation databases, such as GENCODE, Ensembl and UniProt, structure and function information available at the APPRIS webserver and interspecies conservation analysed by PhyloCSF and Compara

Out of the 2,467 potential non-coding genes identified, manual annotators from the HAVANA team have already removed 71 and 28 have been left for discussion.

Keywords

Gene annotation, non-coding genes, GENCODE

OBJETIVOS

El objetivo principal del proyecto es generar un pipeline automatizado que provea una lista de genes etiquetados como potencialmente no codificantes que pueda ser usado por los anotadores manuales de HAVANA para ayudar en la reclasificación de esos genes.

Para ello será preciso definir un set de características potencialmente no codificantes para los genes de humano del *gene set* de GENCODE v27, procedentes de las bases de datos GENCODE, Ensembl, UniProt y APPRIS, así como de los de la puntuación basada en exones de PhyloCSF, que representa una medida de conservación para cada gen, y datos referentes a la edad de la familia génica o a la conservación de especie de ese gen.

El pipeline también deberá ser capaz de anotar las mismas características para el genoma de ratón, con el objetivo último de que sea útil para ayudar en la anotación de otras especies.

Este pipeline se lanzará con cada nueva versión del genoma GENCODE, tanto de ratón como de humano.

INTRODUCCIÓN

El proteoma humano de referencia constituye un pilar fundamentar de la investigación básica en genómica, biología evolutiva y proteómica, y apoya casi todos los proyectos biomédicos de gran escala. Es importante definirlo bien para que no haya consecuencias a nivel de las bases de datos y de los servicios que las usan y para evitar complicaciones en los experimentos biomédicos a gran escala, especialmente aquellos que están relacionados con el mapeo de variaciones relacionadas con enfermedades con genes humanos.

Uno de los primeros intentos de estimar el número de genes corrió a cargo de Friedrich Voguel, cuando el código genético aún no había sido descifrado, hace más de 50 años[1]. Sabiendo que tres aminoácidos corresponden a un nucleótido y basándose en el número de aminoácidos que hay en cada cadena de la hemoglobina, calculó, por extrapolación, el peso del DNA que compone estos genes. Usando el peso molecular del cromosoma humano haploide calculó el tamaño del genoma y dividiéndolo por el tamaño de un gen (el gen de la hemoglobina), llegó a la cifra de 6,7 millones de genes.

En los años previos a la secuenciación del genoma humano la mayoría de investigadores estimaba que el número final de genes codificantes estaría entre 50,000[2] y 80,000[3], con algunos autores incluso apostando por cifras mayores a 100,000[4]. Se esperaba (por parte de muchos genetistas) que la cifra definitiva de genes se resolviera junto con la secuenciación del genoma, pero, en 2001, aún con el 80% del genoma resuelto, seguía siendo una incertidumbre. Este tema fue tan controversial que Ewan Birney, uno de los jefes técnicos y biólogos computacionales del proyecto Ensembl [5] organizó una competición, *Genesweep* [6], en la que cada participante apostaría por una cifra que se resolvería tres años más tarde. Tres grupos de investigación[7,8,9] llegaron a conclusiones similares usando tres métodos independientes (el número de genes codificantes estaba por debajo de 40,000), abriendo el debate sobre la posibilidad de que la clave de la complejidad fisiológica fuera el resultado de la diversificación de redes de regulación o del *splicing* alternativo, no tanto un aumento en el número de genes.

En los años posteriores la tendencia continuó siendo la de disminuir el número total de genes. Michelle Clamp[10] y su equipo en 2007 propusieron una metodología de evaluación de posibles futuras entradas al catálogo de genes definitivos, y bajaron la cuenta de número de genes a 20,500. Unos años más tarde, Church *et al*.[11] llevaron a cabo una comparación entre los genomas humanos y de ratón, para encontrar al igual

que la autora anterior, que la mayoría de los genes codificantes ganados en mamíferos se generan por duplicación, y que realmente el número de genes “nuevos” es muy bajo. Estimaron que el número de genes codificantes sería menor de 20,000.
Estos trabajos constituyen dos de los estudios más exhaustivos del componente codificante del genoma humano, y se llevaron a cabo antes de que otros grupos comenzaran a llevar a cabo una re-anotación sistemática manual de los genes del catálogo de genes humanos.

La anotación del número de genes codificantes de humano comenzó oficialmente con el proyecto Ensembl en 2002[5] cuya primera publicación incluyó 24,000 genes.
Posteriormente surgió el consorcio de GENCODE[12]. Éste nació como un sub-proyecto del proyecto ENCODE (enciclopedia de elementos de DNA). El objetivo de GENCODE es anotar todas las características genéticas basadas en evidencia (genes, tránscritos, secuencias codificantes, etc.) en los genomas humano y de ratón con una gran exactitud. Hay 8 grupos instituciones principales que colaboran con el proyecto: Wellcome Trust Sanger Institute, European Bioinformatics Institute, University of Lausanne, Centre de Regulació Genómica, University of California, Massachusetts Institute of Technology, Yale University y Centro Nacional de Investigaciones Oncológicas. Las anotaciones finales son producto de la unión de anotaciones manuales por parte del grupo de HAVANA y las predicciones generadas por los modelos computacionales de Ensembl.

A pesar de su importancia, el catálogo de genes humanos sigue sin estar finalizado. Los esfuerzos ahora se centran en alcanzar dos objetivos principales: asegurarse de incluir todos los genes codificantes para proteínas y a la vez excluir aquellos que muestran evidencias dudosas. Esta última parte está resultando sorprendentemente difícil para los anotadores manuales.
Hay genes que han sido anotados como codificantes debido a la presencia de evidencia de dominios de proteína que posteriormente resultan ser falsos, experimentos de interacción de proteínas a gran escala que rinden falsos positivos o incluso por evidencia experimental publicada que ha resultado ser dudosa a posteriori.

Una vez que un gen se anota como codificante, se encuentra presente en las bases de datos de gran escala. Cuando un supuesto gen codificante entra en el catálogo de genes humanos, es difícil de eliminar. De hecho, una vez allí, su potencial codificante puede ser validado por anotación circular. La existencia de estos genes en las bases de datos

crea un serio problema, que se sólo se acentúa a medida que los proyectos de secuenciación a gran escala rinden mayores números de transcritos[10].

El mayor desafío, y a la vez el motor de evolución, de las estrategias de anotación es la paulatina compresión de la complejidad transcripcional, ya que los genomas eucariotas no sólo están compuestos por genes codificantes, sino que también comprenden pseudogenes, ARN largo no codificantes (lncRNAs o *long non-coding RNAs* por sus siglas en inglés) y familias de ARN pequeño, que incluyen ARN de transferencia, ARNs asociados a Piwi (piRNAs) y RNA pequeño nucleolar (snoRNAs), e incluso algunos tipos de ARN aún por descubrir.

No hay forma automática o experimental de distinguir genes erróneamente anotados como codificantes. Aunque se puede demostrar la validez de un gen codificante mediante, por ejemplo, evidencia de espectrometría de masas directa de la proteína codificada, el que no se encuentre evidencia de la misma no implica que quede demostrada la invalidez de un cierto gen para codificar proteínas.

Debido a la gran dificultad del proceso de anulación del estatus codificante de un gen que no lo es, un sistema fiable que detecte genes codificantes no válidos, como el implementado en el presente proyecto, es por lo tanto de gran utilidad para los anotadores manuales del genoma humano, extensible también a otras especies.

MATERIALES Y MÉTODOS

Pipeline

Se ha diseñado un pipeline automático haciendo uso del lenguaje de programación Python[13] así como de la línea de comando[14].

El pipeline consta de 20 *scripts* independientes que llevan a cabo el pre-procesamiento de los datos provenientes de las bases de datos, a la vez que ejecutan el análisis necesario para extraer las quince características potencialmente no codificantes, así como la asignación de pesos para cada una de las características.

Formato GTF

Uno de los ficheros resultantes de análisis con el pipeline automático diseñado es un fichero en formato GTF. El formato GTF ("*General Transfer Format*") es idéntico a la versión 2 del formato GFF ("*General Feature Format*"). Ambos consisten en una línea por cada característica, cada una conteniendo 9 columnas de datos, más una línea opcional de definición[5].

Los campos se encuentran separados mediante tabuladores y todos, excepto el último, deben contener un valor (los campos vacíos se denotan con un punto). Los campos se describen a continuación:

A. Columnas estándar delimitadas por tabulador

1. *Seqname*: nombre del cromosoma o del ensamblaje

2. *Source*: nombre del programa que genera la entrada, de la fuente de los datos (base de datos o nombre del proyecto)

3. *Feature*: nombre del tipo de característica que se anota. Por ejemplo, gen o variación.

4. *Start*: Posición de inicio de la característica anotada, comenzando por 1.

5. *End*: Posición de fin de la característica anotada, comenzando por 1.

6. *Score*: Valor decimal. El valor de *score* es la suma de la puntuación de cada característica potencialmente no codificante encontrada para ese gen. '.' indica que no se han encontrado características en ese gen.

7. *Strand*: + ó -.

8. *Frame*: 0, 1 ó 2. 0 indica que la primera base de la característica es la primera base del codon, 1 indica que la segunda base es la primera del codón.

9. *Attribute*: Lista pares etiqueta – valor separada mediante punto y coma, que provee información adicional sobre cada característica.

Número de columna	Contenido	Valores/formato
1	*Chromosome name*	chr{1-22, X, Y, M}
2	*Annotation source*	{ENSEMBL, HAVANA}
3	*Feature type*	Gene
4	*Genomic start location*	Valor entero (1-based)
5	*Genomic end location*	Valor entero (1-based)
6	*Score*[*]	{., 1-7}
7	*Genomic strand*	{+,-}
8	*Genomic phase*	{0,1,2,.}
9	*Additional information*	Ver siguiente sección

B. Pares etiqueta – valor encontrados en la columna 9

Nombre etiqueta	Formato del valor
gene_id	ENSGXXXXXXXXX
gene_type	{IG_X_gene, polymorphic_pseudogene, protein_coding, TR_X_gene}
uniprot_evidence	{Evidence at protein level, Evidence at transcript level, Inferred from homology, Predicted, Uncertain, No evidence}
appris_protein_characteristic	{Yes, No, NA}
uniprot_description	{non-functional, opposite strand, pseudogene, readthrough, protein coding, -}
uniprot_caution	{non protein-coding, pseudogene, none, -}
ensembl_description	{antisense, non-functional, non protein-coding, opposite strand, pseudogene, readthrough, protein-coding, -}
transcript_type	{protein_coding, readthrough/nonsense_mediated_decay}[/]
phyloCSF_score	{above_cutoff, below_cutoff, not_enough_evidence, No score}
gene_family_age	{Fungi-metazoa_group, Bilateria, Chordata, Vertebrata, Euteleostomi, Tetrapoda, Amniota, Mammalia, Theria, Eutheria, Boreoeutheria, Primates, Simiiformes, Catarrhini, Hominoidea, Hominidae, HomoPanGorila, NA}

Tabla 2: Descripción de los campos que componen la 9ª columna del fichero GTF. El formato es etiqueta: "valor".

Bases de datos

A continuación, se encuentra una breve descripción de las bases de datos de las cuales se han extraído los *datasets* de referencia tanto para el genoma humano como para el de ratón.

GENCODE[12]: Proyecto multi-institucional que provee anotación de genes para humano y para ratón, en colaboración con el proyecto ENCODE. Los *genebuilds* son el resultado de modelos manualmente anotados producidos por el grupo HAVANA junto con modelos computacionales generados por Ensembl. Otros grupos proveen validaciones experimentales e *in* silico.

Ensembl: Fuente de anotación genómica que provee anotación de varias especies, así como datos de regulación y de enfermedades. Su anotación genética se basa en un pipeline de análisis propio, que cuenta con anotación manual.
Estas dos bases de datos se unen en una.

UniProt[15]**:** Repositorio de proteínas que incorpora las bases de datos de secuencias de proteínas de Swiss-Prot (anotación manual por expertos) y TrEMBL (entradas analizadas computacionalmente). También reciben anotaciones de GENCODE/Ensembl, pero las entradas no siempre coinciden al 100%.

APPRIS[16]: Ejecuta una variedad de métodos computacionales, basados en estructura de proteínas, funcionalidad y conservación, para seleccionar una isoforma "principal" para cada gen.

Características no codificantes (PNC, *potential non coding*)

Las características en las que se ha basado el proyecto para identificar los genes potencialmente no codificantes se describen a continuación. A cada una de ellas se le ha dotado de un peso determinado, ya que no todas tienen la misma importancia a la hora de determinar la capacidad codificante de un gen. Para cada gen se suma el total de puntuaciones que dan estas características, obteniendo una cifra final según la cual se clasifican como genes probablemente codificantes o genes potencialmente no codificantes.

GENCODE

Elección de transcritos**:** el fichero GTF de anotación de GENCODE incluye para cada gen un número de transcritos. Con el objetivo de escoger un solo transcrito representante de cada gen, se ha hecho uso de la información que GENCODE provee en su anotación, en concreto se han ordenado los transcritos atendiendo a los siguientes criterios, en orden de importancia:

- Tipo de transcrito:
    - Codificante: Anotado como *protein_coding*
    - Genes de la región variable de inmunoglobulina o genes de receptores de células T: Anotados como *IG_X_gene* y *TR_X_gene* (donde X puede ser para ambos C, D, J y V)
    - Pseudogenes polimórficos: anotados como *polymorphic_pseudogene.*
    - Degradación mediada por mutaciones terminadoras: anotados como *nonsense_mediated_decay*
    - Degradación por falta de codón de parada: Anotados como *non_stop_decay*.
- Mejor TSL (*Transcript Support Level*): los transcritos se puntúan de acuerdo a cómo de bien solapan los alineamientos de ARN mensajero y EST (marcador de secuencia expresada) sobre su secuencia.
    - 1: Todos los empalmes del transcrito están respaldados por al menos un ARN mensajero no dudoso.
    - 2: El mejor ARN mensajero está clasificado como dudoso, o está respaldado por multiples ESTs.
    - 3: El único apoyo proviene de un solo EST.
    - 4: El mejor EST está clasificado como dudoso.
    - 5: La estructura modelo no está respaldada por ningún transcrito.
    - NA: el transcrito no ha sido analizado.
- Anotación de APPRIS: en caso de que dos transcritos coincidan en los campos anteriores, se escoge aquel que esté designado como isoforma principal según la base de datos de APPRIS.

Polymorphic pseudogenes*:* GENCODE incluyó la etiqueta de *polymorphic pseudogene* dentro del campo de tipo de gen cuando comenzó a generar un *genset* de genes codificantes. Un pseudogen polimórfico es un pseudogen que no ha sido fijado en el genoma, por lo que coexiste como un alelo junto con el gen funcional en el mismo locus. Algunos genes defectuosos podrían ser perjudiciales, ya que representarían alelos con

pérdida de función que ponen en riesgo la supervivencia del organismo. Estos genes pueden ser ejemplos de pseudogenes que están sustituyendo el gen funcional. Esto indicaría que el gen funcional no se encuentra bajo una selección fuerte. Por ejemplo, un pseudogen procesado generado recientemente puede ser polimórfico en el sitio de inserción y el loci duplicado puede tener alelos que son aún funcionales y otros que son inactivos. La fijación del pseudogen puede tardar bastante tiempo[17].

Se han identificado 63 *polymorphic pseudogenes* en el set de referencia, de los cuales 27 están clasificados como receptores olfativos y 16 no están anotados como tal, aunque por su símbolo HGNC se deduce que son receptores olfativos. De los más de 1,000 genes que son receptores olfativos, cerca de la mitad parecen ser pseudogenes; distinguir cuáles de ellos son pseudogenes y cuáles codifican para proteína es un verdadero reto en el que hoy en día se sigue trabajando[18]. A esta característica se le ha dotado un peso de 2.

Nonsense-mediated decay genes*:* La degradación del ARN mensajero mediada por mutaciones terminadoras (*Nonsense-mediated mRNA decay, NMD*) es un mecanismo de control usado por las células para eliminar ARN mensajero aberrantes que se originan por mutaciones en la línea germinal en muchos desórdenes genéticos, así como los originados por errores durante la transcripción. Recientemente, se ha descubierto que el sistema NMD juega un papel más general en la regulación de la expresión génica controlando la degradación de un subconjunto de ARN mensajero y tiene un papel crucial en la diferenciación de células madre. A pesar de su relevancia, el mecanismo molecular detallado de este proceso no se conoce completamente todavía[19]. GENCODE anota algunos transcritos como NMD y, por razones técnicas, estos pasan a formar parte del *genset* de referencia como proteínas, implicando que hay ciertos genes anotados como codificantes que sólo generan ARN mensajero que van a ser degradado mediante esta vía.

Para identificar estos genes ha sido necesario calcular un porcentaje de transcritos anotados como NMD para cada gen. El número total de transcritos NMD más el número de transcritos "*non stop decay*" se divide entre la suma de transcritos codificantes del gen, anotados como *protein coding*, TR (receptores de células T), IG (genes de la cadena variable de inmunoglobulina), NMD y NSD. Sólo los genes que muestran un 100% de porcentaje son clasificados como "*nonsense mediated decay*", en concreto 220, con un peso de 2.

Readthrough genes*:*

Los genes *readthrough* se generan cuando un transcrito se salta el exón 3' y continua con los exones del siguiente gen (que normalmente es codificante, pero puede ser no codificante o un pseudogen), produciendo proteínas de fusión al traducirse. Aunque es posible que el salto del codón de parada sea una de las formas que tiene las proteínas de ganar nuevos dominios[20], muy pocos transcritos *readthrough* generan proteínas a niveles detectables. Por razones técnicas, estas variantes son anotadas como parte del *geneset* codificante de referencia. Para identificar estos genes llevamos a cabo un cálculo de porcentaje de transcritos *readthrough* de cada gen, quedándonos con aquellos genes en los que todos los transcritos codificantes o NMD (todos los genes NMD son también genes *readthrough*) están etiquetados como *readthrough*. También existen genes que tienen una mezcla de transcritos *readthrough* y codificantes, aunque estos están siendo eliminados gradualmente. En la versión 27 de GENCODE se han identificado 492 genes pertenecientes a esta categoría, con un peso asignado a cada uno de 2.

UniProtKB

Elección de transcritos**:** el fichero de anotación de UniProt incluye para cada gen un número distinto de transcritos. El campo común a la mayoría de bases de datos es el identificador de gen de Ensembl, que UniProt no provee, aunque sí incluye el identificador de transcrito de Ensembl para la mayoría de las entradas. Por lo tanto, es necesario añadir el identificador de gen para cada entrada (basándose en el identificador de transcrito). Una vez hecho esto, se procede a usar la información sobre la anotación proveniente del propio fichero de UniProt para ordenar los transcritos de la siguiente manera:

- Estado: Indica cuándo la entrada ha sido anotada manualmente y revisada por los revisores de UniPRotKB. Las entradas revisadas pertenecen a la sección Swiss-Prot de UniProtKB mientras que las no revisadas pertenecen a la sección TrEMBL de anotación computacional.
- Evidencia: Indica la evidencia de existencia de cierta proteína. Se han definido cinco niveles:
    - Evidencia a nivel de proteína: Identificación clara, experimental o de experimentos a gran escala de interacciones.
    - Evidencia a nivel de transcrito: Existencia de ADN copia.
    - Inferido mediante homología: Una proteína predicha a la que se le ha asignado una familia de proteínas en UniProtKB.

    - Predicha: Proteína predicha a la que aún no se le ha asignado una familia de proteínas.
    - Incierto: Secuencias dudosas, como las que derivan de traducciones de pseudogenes o de ARN no codificante.
- Anotación: La puntuación de anotación de UniProt provee una medida heurística del contenido de la anotación. Las entradas se puntúan bien por presencia o bien por número de ocurrencias. Las anotaciones con evidencia experimental puntúan más alto que las anotaciones predichas o inferidas, de forma que se favorece la curación basada en literatura sobre la anotación automática. La puntuación de una entrada individual es la suma de las puntuaciones de sus anotaciones.
- Advertencia: Informa sobre una gran variedad de posibles errores y/o temas de confusión relevantes a diferentes aspectos de la información provista sobre la proteína.

Evidencia: Las evidencias 'Inferido mediante homología', 'Predicho' y 'Incierto' se consideran como características presentes en los genes potencialmente no codificantes. De un total de 1,528 genes que caen en esta categoría, 624 tiene se han inferido mediante homología, 852 son predichos y los 52 restantes inciertos. Los pesos de cada uno son, respectivamente, 1, 2 y 4.

Advertencia: UniProtKB añade advertencias a muchas de las entradas de proteínas. Algunas de ellas muestran dudas sobre la posible expresión de la misma. No se han seleccionado todas las advertencias disponibles, sino solo aquellas que sugieren que podría ser no codificante, no funcional o un pseudogen. Cabe destacar que UniProt acompaña de una advertencia todas las entradas que tienen como evidencia 'predicho'. De las advertencias utilizadas, mostradas a continuación, todas están presentes en el fichero original, aunque no todas se encuentran una vez se filtra para elegir un solo transcrito por gen. Todas las siguientes advertencias cuentan con un peso de 3:

- 'Could be the product of a pseudogene': Podría ser el producto de un pseudogen. Se han encontrado 40 genes con esta advertencia.
- 'Long intergenic non-protein coding mRNA': ARN mensajero de interferencia.
- 'Probable non-coding RNA': ARN posiblemente no codificante.
- 'May be a non-coding RNA': Podría ser un ARN no codificante. Se ha encontrado un gen con esta advertencia.
- 'May be produced at low levels due to a premature stop codon': Podría producirse a niveles bajos debido a un codón de parada prematuro. Se ha encontrado un gen con esta advertencia.

Nombres de la proteína: Esta sección incluye una lista exhaustiva de todos los nombres de la proteína, desde los más usados hasta los obsoletos, así como información breve sobre la actividad de la proteína (como una descripción del mecanismo catalítico de las enzimas, o información sobre dominios funcionales). Durante el estudio llevado a cabo, se han identificado una serie de palabras que se encuentran a menudo en esta sección y que indican una posible actividad no funcional de la misma. A todas ellas se les ha asignado un peso de 2:

- Cadena complementaria (*opposite strand*): Se han identificado 16 genes con esta descripción.
- *Readthrough*: Se han identificado 68 genes con esta descripción.
- Pseudogen (*pseudogene*): Se han identificado 34 genes con esta descripción.
- No funcional (*non-functional*): Se han identificado 50 genes con esta descripción.

Ausencia de información en UniProt: El hecho de que no haya información en UniProt para un gen ha sido considerado como una característica que indica que el gen puede ser no codificante, por lo que se le ha otorgado un peso de 2. En el *geneset* de referencia se han encontrado 150 genes que no tienen información en UniProtKB.

Ensembl

Elección de transcritos**:** a diferencia de las dos bases de datos anteriores, Ensembl no incluye información acerca de la anotación de las entradas que permita clasificar y ordenar los transcritos para asignar uno por gen. Por lo tanto, ha sido necesario extrapolar la información proveniente del fichero de UniProt en relación a la calidad de la anotación, con la obtenida previamente de BioMart. Una vez añadida la información, se procede a elegir un transcrito para cada gen de forma similar a la descrita anteriormente para la información proveniente de UniProtKB.

Descripción del gen: Ensembl permite descargar un campo denominado descripción del gen, que de forma similar al campo 'Nombres de proteína' de UniProt, proporciona una breve descripción de la funcionalidad del gen. Al igual que en el caso anterior, se han identificado una serie de descriptores que hacen referencia a la falta de funcionalidad del gen, con un peso cada uno de ellos de 2:

- Pseudogen (*pseudogene*): Se han identificado 97 genes con esta descripción, de los cuales 37 son receptores olfativos.
- *Readthrough*: Se han identificado 95 genes con esta descripción.
- No codificante (*non-protein coding)*: Se han identificado 12 genes con esta descripción.

- Anti sentido (*antisense*): Se han identificado 10 genes con esta descripción.
- No funcional (*non-functional*): Se han identificado 55 genes con esta descripción.
- Cadena complementaria (*opposite strand*): Se han identificado 15 genes con esta descripción.

Puntuación de PhyloCSF

Se ha usado una puntuación basada en exones de PhyloCSF[21] para representar medida de conservación de cada gen. La puntuación de conservación corresponde con la puntuación PhyloCSF del exón con puntuación más alta. Los umbrales mínimos establecidos para que la puntuación fuera tomada como válida fueron una longitud de exón de al menos 14 bases y una longitud de rama mínima de 0.075. Los genes que no pasan estos umbrales se consideran como 'sin suficiente información' (*not enough evidence*), considerada como una de las características de genes potencialmente no codificantes, con un peso de 1, al igual que aquellos genes que no cuentan con información disponible de PhyloCSF. De los primeros se han identificado 266, mientras que de los últimos 265.

Los genes con una puntuación de PhyloCSF menor que -15 son marcados como potencialmente no codificantes con un peso de 3. Dentro de esta categoría se han encontrado 101 genes.

Familia génica de primate

Se han utilizado datos de análisis de edad de genes, basados en árboles filogenéticos, como otra característica potencialmente no codificante. Estos datos provienen de un estudio llevado a cabo con anterioridad[22] por lo que corresponden a una versión anterior del *geneset* de referencia de GENCODE, la versión 24.

Para ello, se usaron las reconstrucciones filogenéticas de Ensembl Compara[23], basadas en genes secuenciados de 58 especies diferentes. Se consideran sólo las clases de edades que representan el último ancestro común de *Homo sapiens* y especies secuenciadas con una cobertura relativamente alta (como mínimo 5x). El análisis incluye las siguientes clases de edades para los genes humanos: *Fungi/Metazoa*, *Bilateria*, *Coelomata*, *Chordata*, *Vertebrata*, *Euteleostomi*, *Sarcopterygii*, *Tetrapoda*, *Amniota*, *Mammalia*, *Theria*, *Eutheria* (*Eutheria* + *Euarchontoglires*), *Simiiformes*, *Catarrhini*, *Hominoidea*, *Hominidae*, *HomoPanGorilla* y *Homo sapiens*.

Ensembl Compara clasifica los nodos de especiación y duplicación en árboles teniendo en cuenta el nivel filogenético en el que el evento ocurrió [24]. El análisis se basa en esta información para definir la edad de la familia del gen y la edad del gen para cada gen anotado como codificante. La edad de la familia del gen es el primer ancestro común que tiene un miembro de la familia génica, mientras que la edad del gen es el primer ancestro común en el que tiene lugar el evento genómico que conduce al gen existente. Para los genes duplicados la edad del gen representa la especie en la que sucedió la duplicación, mientras que cuando no hay duplicación (*singleton genes),* la edad de la familia del gen corresponde con la edad del gen.

El que el gen pertenezca a una de los siguientes grupos de edad de la familia se ha considerado como una característica de genes potencialmente no codificantes, con un peso de 3:

- Primates
- Simiiformes
- Catarrhini
- Hominoidea
- Hominidae
- HomoPanGorilla
- Human

El hecho de que no se encuentre información en estos datos también se ha considerado como una característica con un peso de 3. Estos últimos genes son aquellos que se encuentran en los datos provenientes de Compara pero que sin embargo no tienen información asociada. Al pertenecer los datos a una versión del genoma anterior a la actual (v27 de GENCODE), aquellos genes que no se encuentran en la versión actual no han sido penalizados.

APPRIS

Todos los genes de Ensembl están anotados en la base de datos de APPRIS. APPRIS anota las siguientes características de proteína: homología con proteínas de estructura conocida, mapeo de residuos funcionalmente importantes y dominios funcionales de proteínas de *firestar*[25] y pfamscan[26]; las hélices transmembrana son mapeadas usando tres predictores transmembrana independientes[27,28,29], y los péptidos señal se predicen mediante *SignalP*[30]; otro módulo de APPRIS calcula una medida de conservación mapeando ortólogos de vertebrados presentes en la base de datos de proteínas.

Elección de transcritos: Al igual que en los casos anteriores, la base de datos de APPRIS ofrece información sobre todas las variantes anotadas, mientras que para llevar a cabo la comparación con el resto de bases de datos se precisaba una sola variante por gen, por lo que en este caso se escogió la isoforma principal de cada gen.

Características de proteína: Aquellos genes que carecen de información funcional (*firestar* distinto de 0 o SPADE menor de 20), estructural (Matador menor de 0,02 o THUMP menor de 1) o de conservación (CORSAIR menor de 1,5) se han etiquetado como potencialmente no codificantes, con un peso de 3, siempre que su puntuación correspondiente en PhyloCSF sea menor de 2. Dentro de esta categoría se han encontrado 389 genes.

Expresión de transcritos de Human Protein Atlas

Se obtuvo acceso a los datos de experimentos de RNA-seq llevados a cabo por Human Protein Atlas[31]. Estos experimentos se llevaron a cabo en 36 tejidos utilizando la versión de Ensembl v83 (equivalente a GENCODE v24). Para cada gen, se ha contado el número de tejidos en los que los niveles de expresión son como mínimo un transcrito por millón (TPM).

Los genes en los que no hay transcripción se marcan como potencialmente no codificantes, con un peso de 3, encontrando 368 genes.

Los genes para los que no hay información disponible de este experimento no son penalizados ya que, como se acaba de comentar, el experimento se llevó a cabo para una versión anterior a la actual, por lo que es de esperar que el número de genes sin información sea elevado (948).

Datos de proteómica de PeptideAtlas[32]

PeptideAtlas contiene *builds* para organismos individuales y para grupos de muestras importantes (por ejemplo, plasma). Hay un criterio de inclusión de péptidos, de forma que los *builds* con un umbral de “P≥0.9" son aquellos que contienen únicamente péptidos con una probabilidad PeptideProphet de al menos 0.9. Un *build* con un umbral de FDR de PSM (tasa de falso descubrimiento) normalmente implica que es un *build* con un FDR de proteína de 1%. Para estos *builds*, un péptido normalmente debe tener una probabilidad mucho mayor de 0.9.

RESULTADOS Y DISCUSIÓN

Determinación de genes potencialmente no codificantes

En estudios previos que usan una metodología similar[22] se han descrito hasta 2,001 genes con características potencialmente no codificantes, de los que finalmente han sido eliminados 908 del catálogo de genes humanos codificantes.

La versión actual del catálogo de referencia de GENCODE, v27, cuenta con un total de 58,288 genes, de los cuales 20,292 son anotados como codificantes. Hay un total de 200,401 transcritos, 80,930 codificantes. La versión anterior contaba con 20,266 genes codificantes. El catálogo de referencia se encuentra en un estado de cambio continuo en el que se procesan, añaden, eliminan y refinan los genes usando la información proveniente de la anotación manual y automática.

Pipeline

Se ha desarrollado un pipeline automático que, como resultado, genera un fichero de formato GTF (incluido en el browser de UCSC), así como un fichero de datos con información extendida. La descripción del formato GTF se encuentra detallada en el apartado de métodos. El segundo fichero es un fichero de anotación que incluye las características potencialmente no codificantes encontradas para cada gen, y la información que justifica esas características correspondientes a distintas bases de datos. Este fichero de datos se ha dado a los anotadores manuales de GENCODE para comprobar en detalle cada uno de los genes etiquetados como "*potential non-coding*"

Características potencialmente no codificantes

Se ha definido un set de 14 características no codificantes del *geneset* de GENCODE v27. Estas características, detalladas en la sección de métodos, provienen de las bases de datos de GENCODE, Ensembl, UniProtKB y APPRIS, de los datos provenientes de los métodos PhyloCSF y Compara y de datos de RNA-seq provenientes de la base de datos Human Protein Atlas. Como se ha comentado anteriormente, hay estudios anteriores que ya utilizan algunas de estas características para etiquetar genes no codificantes[23].

Las 14 características potencialmente no codificantes, la fuente de la que provienen, el peso del que se dota a cada una, la versión de la que se han obtenido y el número de genes que se han etiquetado con cada una de las características se detallan en la tabla 1.

Característica	Fuente	Peso	Genes
Polymorphic pseudogene	GENCODE	2	63
UniProt Evidence Code:	UniProtKB		1528
Inferred from homology		1	624
Predicted		2	852
Uncertain		4	52
No Evidence Code	UniProtKB	2	150
Lack of Protein Features	APPRIS	3	389
UniProt decription:	UniProtKB	2	168
Opposite strand			16
Readthrough			68
Pseudogene			34
Non-functional			50
UniProt Caution	UniProtKB	3	42
Ensembl description:	Ensembl	2	284
Pseudogene			97
Readthrough			95
Non-protein coding			12
Antisense			10
Non-functional			55
Opposite strand			15
All Transcripts Readthrough	GENCODE	2	492
Nonsense mediated decay	GENCODE	2	220
PhyloCSF poor score	PhyloCSF	3	101
No PhyloCSF information:	PhyloCSF	1	531
Poor alignments			266
No PhyloCSF available			265
Primate Family	Compara	3	215
No Compara information	Compara	3	539
No transcription	RNAseq HPA	3	368

Tabla 1. Listado de características potencialmente no codificantes, junto con pesos de cada una, base de datos de la que provienen y número de genes encontrados en la versión v27 de GENCODE con cada una de las características. Las versiones de todas las bases de datos son coetáneas a la versión 27 de GENCODE, excepto los datos de Compara y de RNA-seq, que son equivalentes a la versión anterior, v24.

Como se puede observar de la tabla 1, las características que cuentan con más genes son las correspondientes a la evidencia de UniProt. De los 2,467 genes potencialmente no codificantes, hay más de 400 que no tienen otro problema más que estar anotados con evidencia de *homology*. De hecho, hay 1538 proteínas que tienen evidencia de UniProt *inferred from homology*, *predicted* o *unclear*. Estos datos concuerdan con el hecho de que la muchos de los genes anotados como codificantes que realmente no lo son provengan de las anotaciones automáticas de Ensembl.
Se han asignado una serie de pesos a cada una de las características debido a que no todas tienen las mismas implicaciones para la funcionalidad del gen. Originalmente cada característica contaba con un indicador, de forma que la puntuación obtenida era la suma de las características presentes de cada gen. Los pesos fueron asignados tras un primer análisis de las estas puntuaciones, para resaltar los casos más acusados y facilitar de esta forma la posterior revisión de los anotadores manuales. La característica que cuenta con un peso más elevado es la evidencia de UniProtKB de *uncertain*, ya que implica que no se ha encontrado homología con ninguna familia de proteínas y que además probablemente procede de una traducción de pseudogenes o de ARN no codificante. Como se comenta en la sección de Materiales y Métodos, hay características que en un primer análisis se habían considerado, y a las que se les ha asignado posteriormente un peso de 0, principalmente por la imposibilidad de incluir genes que sólo aparecen en la nueva versión del genoma, al tratarse de datos provenientes de versiones anteriores (Compara y RNA-seq).

La versión actual de GENCODE, v27, cuenta con un total de 20,292 genes codificantes, de los cuales 2,467 cuentan con al menos una de las 15 características.

Hay 961 genes que cuentan con una puntuación mayor de 5, y sólo uno de ellos cuenta con evidencia de péptidos, CDRT15 (*CMT1A duplicated region transcript 15 protein*), que tiene una puntuación de 6 (no tiene característica de proteína según APPRIS y la edad de su familia génica es reciente) y sólo un péptido encontrado. Constituye un buen ejemplo de la dificultad a la que se enfrentan los anotadores a la hora de clasificar este tipo de genes. Este gen proviene de una duplicación de CMT1A, en el cromosoma 17. Comparte la posición de inicio y de fin de CDS con CDRT15L2 (anotado también como potencialmente no codificante). Está en nuestra lista porque no tiene puntuación de conservación alta, ni cuenta con ningún dominio Pfam. La única publicación en la que se nombra este gen [33] argumenta que no se encuentra expresado en tejido adulto, pero sí en tejido embriónico. Parece que el gen está anotado como codificante este caso

porque, tal y como se observa en la ilustración 1, posteriormente a esa publicación se ha encontrado con bastante evidencia en experimentos de proteómica en testículo.

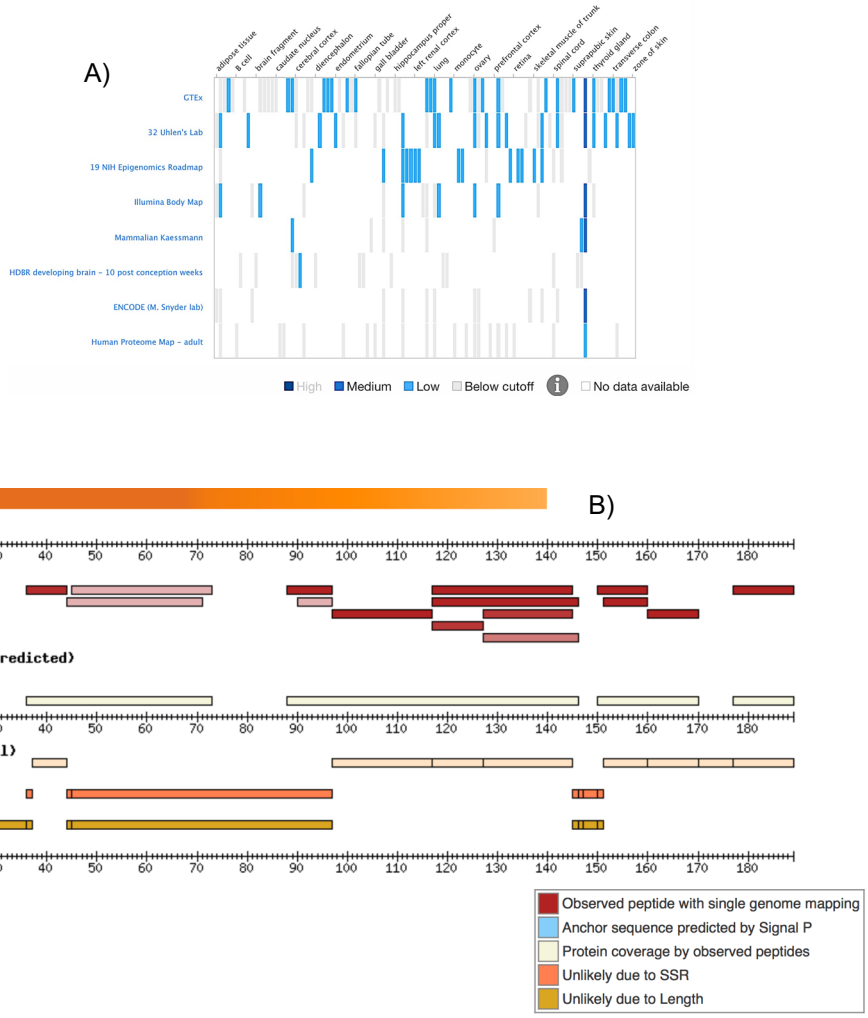

Ilustración 1. Datos de expresión de CDRT15 procedentes de Expression Atlas (A) y Peptide Atlas (B).

A) Expression Atlas representa la expresión de los genes mediante heatmaps. Así, traduciendo los valores de expresión a códigos de colores, se genera una representación visual de los niveles de expresión de los genes sometidos a diferentes condiciones biológicas o procedentes de distintos experimentos. En esta imagen se aprecia cómo el gen se encuentra expresado principalmente en testis, habiendo sido validado por varios experimentos independientes.

B) Encontramos un diagrama gráfico, que resume todos los péptidos que mapean o la proteína, así como información sobre segmentos con poca probabilidad de ser observados con espectrometría de masa, así como péptidos señal o dominios transmembrana. Los péptidos observados se colorean en rojo en la secuencia de la proteína. Podemos observar cómo para CDTR15, los péptidos observados cubren el 83,9% de la secuencia observable.

Se ha descrito que los genes no codificantes tienen niveles de expresión muy inferiores a los genes codificantes.[34] Es por esto que, aparte de hacer uso de los datos de RNA-seq de Human Protein Atlas para describir una característica no codificante, se han utilizado para llevar a cabo una comparación entre el set de genes potencialmente no

codificantes (eliminando aquellos etiquetados únicamente por falta de datos de RNA-seq) y el set de genes a los que no se les ha etiquetado con ninguna de las características descritas previamente, con el objetivo de ver si se comportan de forma diferentes.

Se unieron los genes por número de tejidos en los que se detecta al menos 1 TPM para comparar las distribuciones de ambos conjuntos de genes. Los resultados se muestran en la ilustración 2.

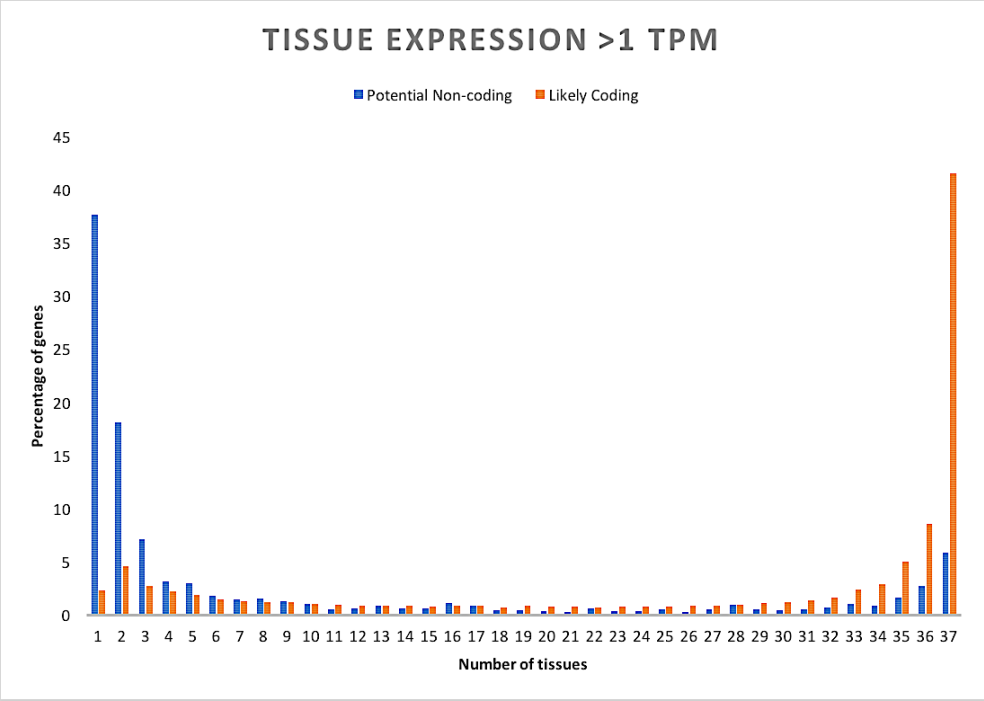

Ilustración 2. Expresión por tejidos del set de genes potencialmente no codificantes y el set de genes codificantes. *Se agrupan los genes por número de tejidos en los que la expresión es igual o superior a 1 TPM en los 36 tejidos de RNA-seq procedentes de Huma Protein Atlas (eje X). El eje Y muestra el porcentaje de genes de cada set. La distribución se muestra para los 1,596 genes potencialmente no codificantes y para los 17,747 genes codificantes.*

La figura muestra dos distribuciones notablemente distintas. El hecho de que muchos genes potencialmente no codificantes tengan niveles de expresión de RNA menores (37% no tienen expresión detectable en ningún tejido) apoya la hipótesis de que muchos de los genes de este *set* no codifican proteínas.

Mientras que los genes codificantes tienden a expresarse en cantidades detectables en la mayoría de tejidos (65,6% de estos genes se detectan en 28 de los 36 tejidos), la mayor parte de los genes potencialmente no codificantes se encuentran en pocos tejidos (el 70,7% de estos genes tienen expresión detectable en menos de 5 tejidos).

Se han seleccionado péptidos identificados de la última versión de PeptideAtlas[32] (enero de 2018). PeptideAtlas forma parte del Human Proteome Project[35], el consorcio que intenta identificar evidencia de proteína para todos los genes codificantes. Según este consorcio, para validar un gen como codificante tiene que tener al menos dos péptidos de tamaño 9 que no solapen entre sí.
Siguiendo este criterio, se han detectado péptidos para 15,252 de los 17,740 genes codificantes (un 85.975%) y 374 de los 2,467 genes con al menos una característica no codificante (15.16%). Es decir, sólo un 2% de la totalidad de los genes identificados por PeptideAtlas son genes potencialmente no codificantes.

Resultados procedentes de HAVANA

Tal y como se ha comentado anteriormente (en la sección de introducción), el equipo de HAVANA se encarga de la anotación manual de los genes de Ensembl, y por lo tanto también de GENCODE. No solo anotan genes codificantes, también ponen especial énfasis en los pseudogenes, los genes *non-coding* y en las variantes de *splicing* alternativo, dos áreas que aún no se han desarrollado en los sistemas de anotación automática.

El objetivo final de este proyecto es que los resultados sirvan de guía a los anotadores manuales a la hora de ver qué conjunto de genes han de revisar para las futuras versiones. En este sentido, la lista con los 2,467 genes potencialmente no codificantes fue puesta en conocimiento del grupo HAVANA. Teniendo en cuenta que el procesamiento que requieren a partir de ahí estos genes es totalmente manual, en la fecha de redacción de este trabajo se habían revisado un total de 467 genes. 65 de ellos han sido eliminados de cara a la próxima versión, mientras que 382 se mantienen.
Dentro de este último grupo, cabe destacar que hay 255 genes que pertenecen al grupo de receptores olfativos, que han sido sujetos a una curación manual detallada. También hay 77 genes que pertenecen a la familia KRTAP (*keratin associated protein*), conocidos por ser altamente dinámicos, difíciles de procesar de forma automática (muy pequeños y repetitivos) y que tienen una expresión bastante restringida. Esto deja un total de 50 genes que se mantienen como codificantes debido a razones funcionales o de conservación (entre su propia familia génica o con especies cercanas).
Hay 20 genes que se han quedado para discusión (a pesar de la evidencia en contra de que sean codificantes), bien porque se encuentren de forma frecuente en la literatura, o porque se encuentran evidencias experimentales (basándose principalmente en datos de proteómica de Human Protein Atlas) o de conservación.

Por ejemplo, el gen HMHB1 es el precursor del antígeno de histocompatibilidad HB-1. Grosso modo, los antígenos menores de histocompatibilidad envían péptidos inmunogénicos que, cuando forman un complejo con MHC, pueden generar una respuesta inmune tras el reconocimiento específico de células T. Este antígeno se encuentra expresado en ciertas células tumorales (Ilustración 1). Muestra tres características no codificantes relacionadas con la conservación (puntuación de PhyloCSF baja, no tiene características de proteína según APPRIS y la edad de la familia génica es relativamente reciente).

Tiene bastante evidencia en la literatura, por lo que la discusión se centra en si debería anotarse como un transcrito aberrante en vez de como una proteína real.

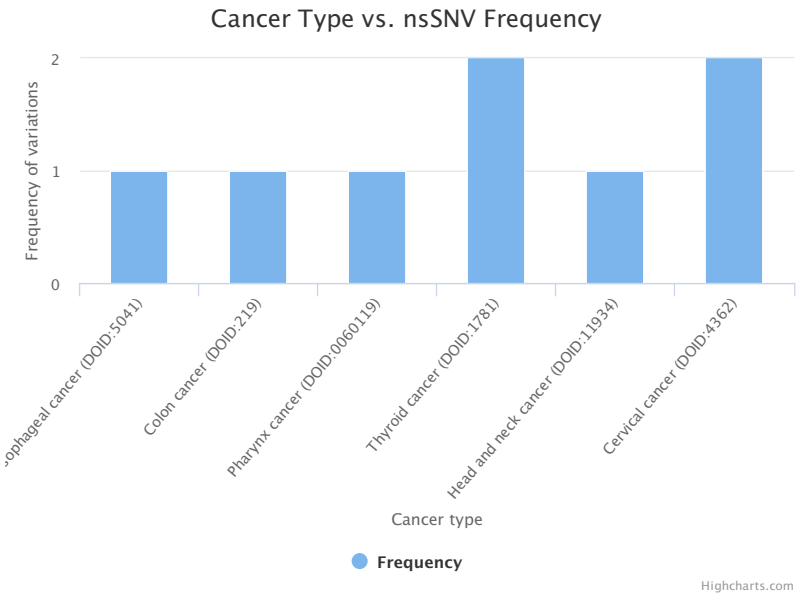

Ilustración 3. Imagen proveniente de BioMuta [36]

Muestra la frecuencia de nsSNVs del gen HMHB1 (eje Y) en el geneset para cada cáncer con el que se le ha asociado (eje X). Las barras más altas indican mayor frecuencia de variantes en el gen para el tipo de cáncer correspondiente.

Otro ejemplo es el gen UBE2Q2L (*ubiquitin conjugating enzyme E2 Q2 like*). Las enzimas de conjugación de ubiquitina, también conocidas como enzimas E2, llevan a cabo el segundo paso de la reacción de ubiquitinación que marca las proteínas para su degradación mediante el proteosoma. El proceso de ubiquitinación une covalentemente ubiquitina, una pequeña proteína de 76 amino ácidos, a un residuo de lisina de la proteína diana[37].

Este gen se encuentra anotado como codificante por GENCODE, aunque UniProtKB le asigna una evidencia de *uncertain* y añade precauciones.

Llevando a cabo un alineamiento del gen con el gen padre (UBE2Q2, ilustración 4) se observa cómo la secuencia se encuentra conservada. Sin embargo, la secuencia hija carece del dominio PFAM de unión a ubiquitina.

El hecho de que la secuencia se encuentre conservada (a pesar de perder gran parte de la secuencia del gen padre) indica que no ha adquirido ninguna nueva funcionalidad, y teniendo en cuenta que no posee la cisteína en posición 305 (que lleva a cabo la unión covalente), tampoco conserva la funcionalidad de la proteína del gen padre, por lo que lo más probable es que se trate de un pseudogen.

Ilustración 4. Proveniente del alineamiento de las secuencias de los genes UBE2Q2 (superior) y UBE2Q2L (inferior) en UniProt.

El residuo señalado en rojo corresponde al residuo funcional anotado por UniProt, mientras que las zonas sombreadas corresponden a aminoácidos con propiedades similares.

DNAH10OS (*dynein axonemal heavy chain 10 opposite strand*, también conocido como *disrupted in renal carcinoma 1*, ilustración 5) codificaría una proteína de 163 amino ácidos cuyo CDS (*coding sequence*) fue predicha en 2009 en base a un análisis completo de transcriptoma., supuestamente con evidencia proteómica. Sin embargo, no hay evidencias funcionales ni otros datos experimentales, o de conservación disponibles y no se encuentra como codificante en la base de datos de RefSeq, por lo que ha sido convertido a *antisense* (como su nombre indicaba).

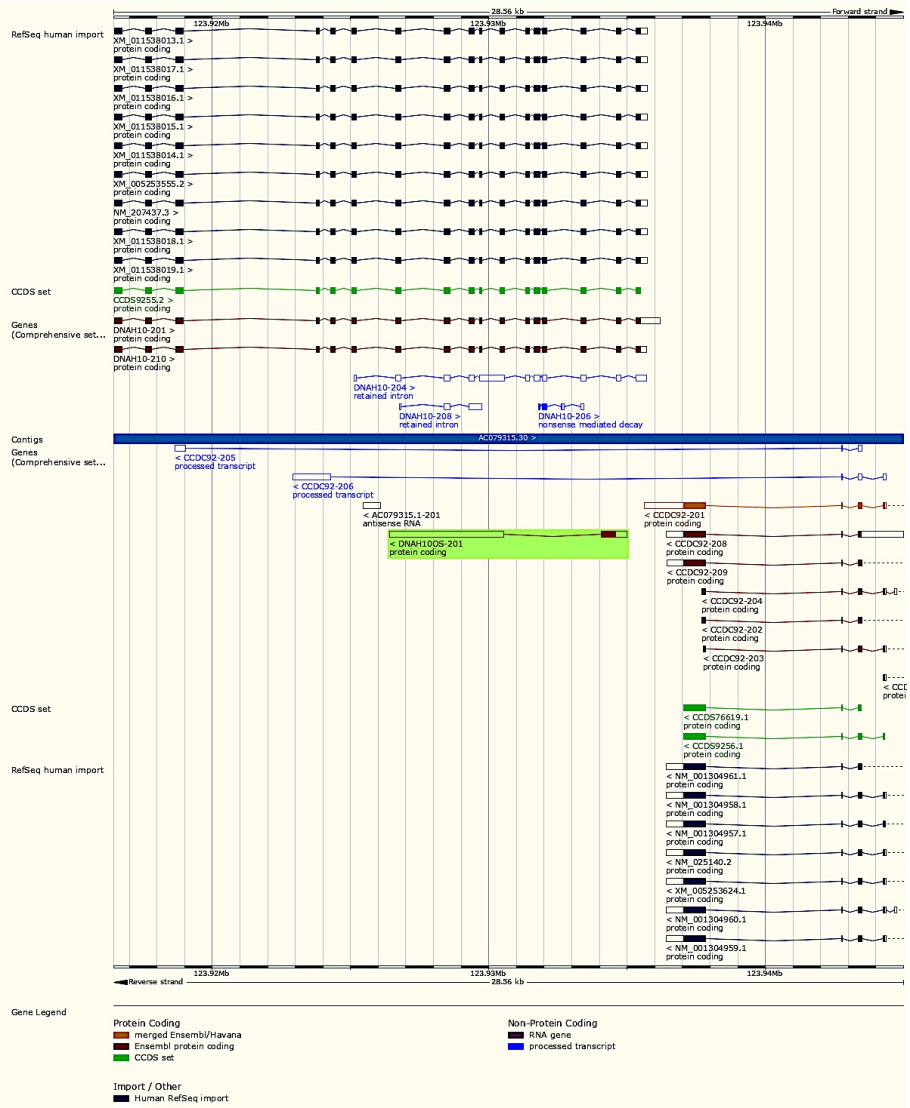

Ilustración 5. Diagrama del gen DNAH10OS procedente de su entrada en Ensembl.

Se observa cómo el gen se encuentra en la hebra opuesta y cómo no hay correspondencia de regiones codificantes con RefSeq ni con CCDS.

DRICH1 (*aspartate rich protein 1*, ilustración 6) constituye otro ejemplo de discusión, ya que no muestra ninguna evidencia de conservación, pero parece tener evidencia a nivel de proteína por parte de anticuerpos policlonales de HPA, Bgee (una base de datos que recopila y compara patrones de expresión génica de distintos tipos de datos, como RNA-seq, Affymetric, hibridación *in situ* y datos EST[38]) y GeneVisible[39].

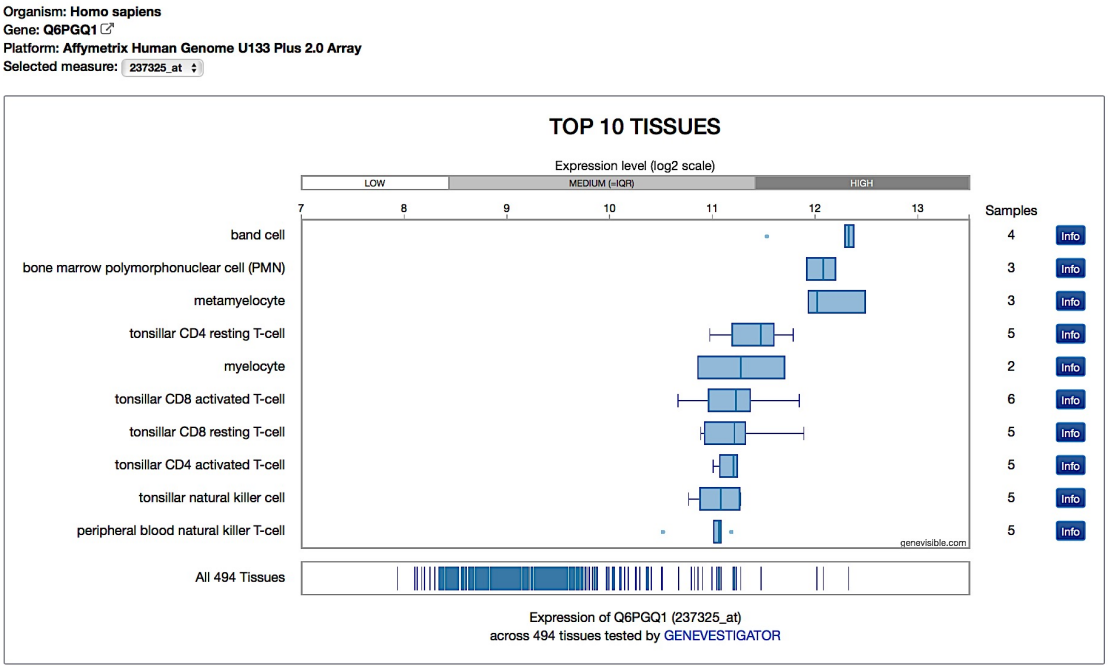

Ilustración 6. Imagen procedente de GeneVisible

Se muestra el top 10 de tejidos en los que se encuentra expresado el gen DRICH1. El más expresado corresponde a granulocitos

Resultados obtenidos del genoma de ratón

Se ha llevado a cabo un análisis preliminar con los datos procedentes de genoma de ratón, a la espera de obtener los datos de Compara y de RNA-seq.

De un total de 22,521 genes codificantes, 1,417 tienen un score de al menos 1.

Los dos genes que mayor puntuación tienen (un 8) son **Trpc5os** (*transient receptor potential cation channel, subfamily C, member 5, opposite strand*) y **Phxr2** (*Putative per-hexamer repeat protein* 2). La distribución de puntuaciones se encuentra representada en la ilustración 7.

Algunas diferencias con respecto a los resultados en humano son sorprendentes. Sólo hay un gen con evidencia de anotación de UniProt *uncertain*, mientras que en humano encontramos 52, y muy pocos genes *predicted* (36 frente a 852). Como contraste, hay más de 7,000 genes de Ensembl que no tienen entrada en UniProt. Es por esto por lo que se decidió modificar el sistema de puntuaciones y no penalizar la ausencia de información de UniProt (de manera contraria, más del 90% de los genes anotados por el pipeline como potencialmente no codificantes no están en UniProt, mientras que con el nuevo sistema el porcentaje se reduce a 66%). Como resultado de lo poco avanzada que está la anotación de UniProt para ratón, hay muy pocos genes que cuenten con una descripción o *caution* que implique que tenga características no codificantes, ya que aún no se han mirado a fondo estos genes. También sorprende que hay pocos genes con baja puntuación de PhyloCSF.

Aparte de la anotación de PhyloCSF, esto concuerda con una anotación bastante menos curada manualmente. También indica que UniProt no incluye automáticamente las anotaciones de genes de ratón procedentes de Ensembl en su proteoma.

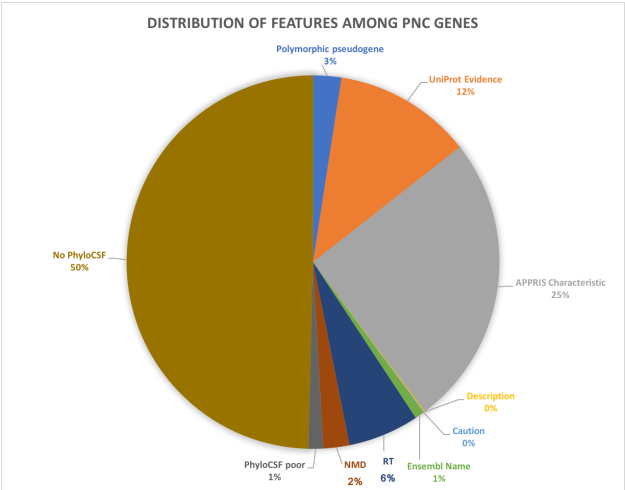

Ilustración 7. Distribución de características potencialmente no codificantes en el set de genes con puntuación > 1. *Podemos apreciar cómo el conjunto de genes que carecen de información de PhyloCSF es el mayor. El conjunto de genes que no poseen características de proteínas según APPRIS es el segundo más numeroso. El porcentaje de genes que se etiquetan por los cautions o por su nombre de UniProt es insignificante en comparación con el resto de genes.*

CONCLUSIONES

Se ha llevado a cabo una revisión del proteoma humano haciendo uso de datos procedentes las tres bases de datos principales de proteomas de referencia. En base a 14 características potencialmente no codificantes se han identificado 2,467 genes potencialmente no codificantes, que corresponden al 12% del total (20,292 genes anotados).

Hay un tipo de genes que merecen una atención especial debido a las dificultades técnicas que presenta su clasificación y su distinción de los genes codificantes. Los pseudogenes son difíciles de separar de genes codificantes cuando son el resultado de una duplicación reciente porque habrán acumulado pocas mutaciones deletéreas. Los *unitary pseudogenes*, un tipo especial de pseudogenes, son codificantes en algunos animales, pero no en humanos (no son procesados y no tienen homólogos funcionales; se generan mediante mutaciones dañinas que ocurren en genes funcionales). Al igual que los genes codificantes de los que provienen, tienen características de proteínas, generalmente están conservados, provienen de familias antiguas y normalmente se transcriben. Esto significa que la probabilidad de ser captados por los métodos utilizados en el *pipeline* es baja.

Sin embargo, parece haber una clara diferencia a nivel celular, ya que los genes marcados como pseudogenes parecen tener una evidencia más clara de selección neutral[41]. Por lo tanto, haciendo uso de estudios de variación genética, se podría analizar el tipo de evolución bajo la que se encuentran estos genes. Por ejemplo, se podría comprobar el número de CNV (*copy number variation)* y la razón de sustituciones no sinónimas y sinónimas (K_a/K_s ratio).

Inevitablemente, la mayor parte del trabajo se ha centrado en la anotación del genoma humano, y muchos aspectos de la anotación de genomas se explican de forma más efectiva en este contexto. Sin embargo, aunque los flujos de trabajo de anotación del genoma humano frecuentemente se aplican en la descripción de otros genomas, estos proyectos no son realmente análogos[41]. Aun así, los genomas de humano y de ratón coinciden en la procedencia de anotaciones que son generadas de forma independiente, como los *genebuilds* de RefSeq y GENCODE.

En base a los resultados preliminares del análisis en ratón, podemos decir que se han identificado un 6% de genes que cuentan con características que los marcan como candidatos a no codificantes. Estos resultados demuestran que el potencial de esta herramienta reside principalmente en su posible uso en la anotación de diferentes especies, donde los anotadores manuales se enfrentan a más dificultades a la hora de

encontrar datos que les permitan contrastar o dotar la anotación de evidencia. Sin embargo, esta misma falta de recursos es la que podría obstaculizar el uso de este sistema de identificación de características potencialmente no codificantes en algunas especies.

Este análisis ha descubierto algunos casos interesantes. Por ejemplo, en 2009 se publicó un artículo[42] según el cual habían encontrado 3 genes presentes en el genoma humano que no lo estaban en el de chimpancé (DNAH10OS, CLLU1 y C22orf45), es decir, tres genes que surgidos de nuevo en humano. Los tres genes fueron anotados como codificantes a raíz de este estudio. En 2013, uno de los genes fue renombrado como ADORA2A-AS1, y clasificado como *antisense* lncRNA.
Tras el análisis llevado a cabo en este proyecto, los otros dos genes, DNAH10OS (comentado en el apartado de resultados) y CLLU1, fueron clasificados como potencialmente no codificantes, y han sido eliminados por el equipo de anotadores de HAVANA recientemente. El articulo llegó a predecir unos 25 genes codificantes humanos que han evolucionado desde que nos separamos de chimpancés. En cambio, nuestros resultados sugieren que no hay ningún gen codificante nuevo en humano.

Este pipeline se va a correr para cada versión nueva de las anotaciones de GENCODE para humano y ratón. A partir de los resultados, se va a poder refinar la anotación del genoma humano con el fin de llegar a una versión final y estable de los sets de referencia de los genes codificantes de las dos especies.

El análisis muestra la importancia de un set de referencia humano revisado. Con el paso de los años, desde que la secuencia del genoma humano se hizo pública, las anotaciones manuales han permitido que nos acerquemos a una versión relativamente final del catálogo de genes humanos codificantes. Sin embargo, aún queda un porcentaje de genes más difíciles de clasificar, debido a las evidencias contradictorias que se muestran. Los datos de estudios a gran escala de variación genética podrían convertirse en una herramienta bastante útil a la hora distinguir los genes codificantes de pseudogenes o de genes no codificantes.

1. VOGEL, F. A Preliminary Estimate of the Number of Human Genes. *Nature* **201,** 847–847 (1964).
2. Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nat. Genet.* **7,** 345–346 (1994).
3. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 11995–9 (1993).
4. Quackenbush, J. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25,** 239–240 (2000).
5. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).
6. No Title. Available at: http://www.ensembl.org/genesweep.html.
7. Sakaki, Y. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
8. Crollius, H. R. *et al.* Estimate of human gene number provided by genome- wide analysis using. *Nat. Genet.* **25,** 235–238 (2000).
9. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25,** 232–234 (2000).
10. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci.* **104,** 19428–19433 (2007).
11. Church, D. M. *et al.* Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. *PLoS Biol.* **7,** e1000112 (2009).
12. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22,** 1760–1774 (2012).
13. Python Software Foundation , Python.
14. Free Software Foundation . Bash (3.2.48) [Unix shell program]. (2007). Available at: http://ftp.gnu.org/gnu/bash/bash-3.2.48.tar.gz.
15. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).
16. Rodriguez, J. M. *et al.* APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx997
17. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13,** R51 (2012).
18. Malnic, B., Gonzalez-Kristeller, D. C. & Gutiyama, L. M. *Odorant Receptors*. *The Neurobiology of Olfaction* (CRC Press/Taylor & Francis, 2010).
19. Llorca, O. Structural insights into nonsense-mediated mRNA decay (NMD) by electron microscopy. *Curr. Opin. Struct. Biol.* **23,** 161–167 (2013).

20. Buljan, M., Frankish, A. & Bateman, A. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.* **11,** R74 (2010).
21. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** i275–i282 (2011).
22. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* **23,** 5866–5878 (2014).
23. Flicek, P. *et al.* Ensembl 2011. *Nucleic Acids Res.* **39,** D800–D806 (2011).
24. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19,** 327–335 (2009).
25. Lopez, G., Maietta, P., Rodriguez, J. M., Valencia, A. & Tress, M. L. firestar — advances in the prediction of functionally important residues. *Nucleic Acids Res.* **39,** W235–W241 (2011).
26. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40,** D290–D301 (2012).
27. Viklund, H. & Elofsson, A. Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.* **13,** 1908–1917 (2004).
28. Käll, L., Krogh, A. & Sonnhammer, E. L. . A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* **338,** 1027–1036 (2004).
29. Jones, D. T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **23,** 538–544 (2007).
30. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2,** 953–971 (2007).
31. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347,** 1260419–1260419 (2015).
32. Desiere, F. *et al.* The PeptideAtlas project. *Nucleic Acids Res.* **34,** D655–D658 (2006).
33. Inoue, K. *et al.* The 1.4-Mb CMT1A duplication/HNPP deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes. *Genome Res.* **11,** 1018–33 (2001).
34. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–89 (2012).
35. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data

Interpretation Guidelines 2.1. *J. Proteome Res.* **15,** 3961–3970 (2016).

36. Wu, T.-J. *et al.* BioMuta. A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). *Database (Oxford).* **2014,** bau022 (2014).
37. Valimberti, I., Tiberti, M., Lambrughi, M., Sarcevic, B. & Papaleo, E. E2 superfamily of ubiquitin-conjugating enzymes: constitutively active or activated through phosphorylation in the catalytic cleft. *Sci. Rep.* **5,** 14849 (2015).
38. Bastian, F. *et al.* in *Data Integration in the Life Sciences* 124–131 (Springer Berlin Heidelberg, 2008). doi:10.1007/978-3-540-69828-9_12
39. Grennan, A. K. Genevestigator. Facilitating Web-Based Gene-Expression Analysis. *PLANT Physiol.* **141,** 1164–1166 (2006).
40. Li, W.-H., Gojobori, T. & Nei, M. Pseudogenes as a paradigm of neutral evolution. *Nature* **292,** 237–239 (1981).
41. Mudge, J. M. & Harrow, J. The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.* **17,** 758–772 (2016).
42. Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res.* **19,** 1752–9 (2009).